AD-A047 931 UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES DEPT O--ETC F/G 5/10
EVALUATIONS OF IMPLIED ORDERS AS A BASIS FOR TAILORED TESTING U--ETC(U)
SEP 77 N CLIFF, R CUDECK, D MCCORMICK N00014-75-C-0684
UNCLASSIFIED TR-4 NL

1 OF 1
ADA047931

END
DATE FILMED
1-78
DDC

12

# EVALUATIONS OF IMPLIED ORDERS AS A BASIS FOR TAILORED TESTING USING SIMULATIONS

Norman Cliff
Robert Cudeck
Douglas McCormick

Technical Report No. 4

TR-4

Department of Psychology
University of Southern California
Los Angeles, California 90007

Sept 1977

62 p.

RR042P4

RR042P401



Prepared under contract No. N00014-75-C-0684
NR No. 150-373, with the Personnel and
Training Research Programs, Psychological Sciences Division

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

| | |
|---|---|
| 1. REPORT NUMBER: Technical Report No. 4 | 2. GOVT ACCESSION NO. |
| 3. RECIPIENT'S CATALOG NUMBER | |
| 4. TITLE *(and Subtitle)*: EVALUATIONS OF IMPLIED ORDERS AS A BASIS FOR TAILORED TESTING USING SIMULATIONS | 5. TYPE OF REPORT & PERIOD COVERED: Technical |
| | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s): Norman Cliff, Robert Cudeck and Douglas McCormick | 8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0684 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Psychology, University of Southern California, Los Angeles, California 90007 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61153N, RR042 04, RR042 04 01, NR 150-373 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Program, Office of Naval Research (Code 458), Arlington, Virginia 22217 | 12. REPORT DATE: September, 1977 |
| | 13. NUMBER OF PAGES: 58 |
| 14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)* | 15. SECURITY CLASS. *(of this report)*: Unclassified |
| | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

(U) Testing; (U) Computer Interaction; (U) True Score; (U) Implied Orders

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

TAILOR is a computer program that uses the implied orders concept as the basis for computerized adaptive testing. The basic characteristics of TAILOR, which does not involve pretesting, are reviewed here and two studies of it are reported. One is a Monte Carlo simulation based on the Birnbaum model and the other nuses a matrix of item responses to the Stanford-Binet.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

In the Birnbaum model study, a variety of conditions were simulated and it was found that TAILOR typically used responses to about half the items and achieved validities with true score within a few points of the validity of the complete test. Item discrimination parameters affected the efficiency of TAILOR.

The Binet study used correlations between scores based on one bank of Tailored items and another independent, parallel set and found results similar to those in the Birnbaum simulation. It appears that TAILOR, like other adaptive testing systems, can aid efficiency where item discriminations are high or, equivalently, ability variance is large.

Contents

# Evaluations of Implied Orders as a Basis for Tailored Testing Using Simulations

Norman Cliff, Robert Cudeck, and Douglas McCormick

The basic principle of the Implied Orders system is a simple one. It arises from considering dichotomous items as furnishing ordering relations between persons and items. If the relations are consistent with each other, then taken as a whole they furnish a joint ordering of the persons and the items. It is well known that the logical properties of an order are such that if certain of the relations among elements are known, then the remainder can be deduced from them by making use of the transitivity property which characterizes orders. This is the principle that underlines our approach to computer-interactive testing. Its basis is spelled out in a 1975 article (Cliff, 1975). The general idea is that even an incomplete matrix of responses of persons to items can be used to deduce at least some order relations between items, which is their relative difficulty. These order relations between items in turn can be used to predict what the individual's responses will be to items not yet taken, and therefore re-

move the necessity of asking those items.

Taking a joint order as a model for test items is equivalent to assuming that the data provide a Guttman scale, but everyone, even we, knows that test data are not Guttman scales. The idea however, is that the joint order is an approximate model for such data. Then the problem in tailored testing is one of how to modify the transitivity principle to make it work reasonably well in the presence of error. The approach that is used here is a rough and ready statistical one. At any given time, there are a certain number of responses that imply that item j is harder than k, and a certain number that imply the reverse. If one kind predominates over the other, then it is implied that one item is easier than the other. Similarly, the pattern of responses to date by an individual may be such that a certain number of them imply that he should get a particular item, which is as yet untaken, wrong; correspondingly, some others may imply that he should get it right. If one number predominates over the other, then the implication is made correspondingly.

## Procedure

### General Idea

Table 1 provides an illustration of the way the procedure operates. The two columns on the left show the responses of 15 persons to two items j and k. To determine which of two items is easier, we examine $n_{jk}$, the number who get j right and k wrong, in comparison to $n_{kj}$, the number who do the reverse. In the data illustrated, person 5 is the only one who gets j right and k wrong, whereas 4, 6, 7, 8, 9, 10 and 11 do the reverse. The frequencies 1 and 7 ($n_{kj}$ and $n_{jk}$ respectively) are then shown at the

Table 1

Illustrative Basis of TAILOR Process

| Complete | | | | Incomplete | | | |
|---|---|---|---|---|---|---|---|
| | Items | | Dominant | | Items | | Dominant |
| Persons | j | k | Item | Persons | j | k | Item |
| 1 | 1 | 1 | -- | 1 | 1 | 1 | -- |
| 2 | 1 | 1 | -- | 2 | 1 | | -- |
| 3 | 1 | 1 | -- | 3 | | 1 | -- |
| 4 | 0 | 1 | j | 4 | 0 | | -- |
| 5 | 1 | 0 | k | 5 | | 0 | -- |
| 6 | 0 | 1 | j | 6 | | | -- |
| 7 | 0 | 1 | j | 7 | 0 | 1 | j |
| 8 | 0 | 1 | j | 8 | 0 | 1 | j |
| 9 | 0 | 1 | j | 9 | 0 | | -- |
| 10 | 0 | 1 | j | 10 | | | -- |
| 11 | 0 | 1 | j | 11 | 0 | 1 | j |
| 12 | 0 | 0 | -- | 12 | | | -- |
| 13 | 0 | 0 | -- | 13 | | 0 | -- |
| 14 | 0 | 0 | -- | 14 | 0 | | -- |
| 15 | 0 | 0 | -- | 15 | 0 | 0 | -- |

| Dominance Frequencies | |
|---|---|
| j | k |
| 7 | 1 |

p = 9/256,
therefore j > k

| Dominance Frequencies | |
|---|---|
| j | k |
| 3 | 0 |

p = 1/8 ,
therefore j > k

bottom of the table. Use of a statistical decision rule, which is outlined below, would lead to the decision that k was easier than j. For each pair of items in a test, such a comparison is made by means of the decision rule. The results of the comparisons are recorded in what we call the item dominance matrix. In the matrix, a 1 means that the row item is harder than the column item.

The foregoing is applicable to a complete test. In an incomplete or tailored test, some of the responses would be missing, as shown in the two righthand columns of the table. The quantities $n_{jk}$ and $n_{kj}$ can still be counted, however. Since person 5 has now only one item, $n_{kj} = 0$. Of the seven persons who had j wrong and k right, we now have data on both items from only three, persons 7, 8 and 11, so $n_{jk}$ is three. The comparison of $n_{jk}$ to $n_{kj}$ could still lead to the conclusion that j was harder than k if this was all the information that was available, provided that the liberal rule were used.

Now consider person 2. He got the harder item right, and has not taken the easier one. We could conclude that he would get the latter right also, and would not give it to him. Similarly, person 13 has the easier item wrong, so we could conclude that he would get the harder one wrong also if he were to take it, and therefore not bother to give it to him. Actually, what we do in making decisions of this kind is similar to what is done with respect to deciding which items are easier and which harder. Suppose person i has not yet taken item j. We look at the number of harder items he has passed and compare that to the number of easier items he has failed. If, by the same decision rule used earlier, the latter of these preponderates over the other, he corres-

pondingly is implied to have failed item i also; if the reverse, then he is assumed to have passed it. If neither preponderates, then no decision is made, and there is no implied response by the person to the item.

Specific Operations

In the program description (Cudeck, et al, Note 1), the frequency-comparing routines above are part of subroutines SQUARE, which computes item dominances, and MULT, which computes implied item responses. These are the major uses, but the comparisons are used in two other places as well. The binary item dominance matrix resulting from SQUARE is multiplied by itself to derive "second order" item dominances. That is suppose, not enough persons have taken both items j and p to yield an implied difficulty order between them, but j has been found to be harder than k and k is harder than p. This suggests that j is harder than p. Evidence of this kind, comparing j to p through all the other items, is used with the significance test in just the same way that it is in the others. This is done in the subroutine called IMPLY. There is one final place where it is used, and this is to help define the order of the persons using the subroutine COUNT. Here, the implied rights matrix is multiplied by the implied wrongs (each resulting from MULT) in order to see if person i has significantly more items right that person h gets wrong than the reverse. If so, i ranks above h. Thus the frequency-comparison process is used to establish direct difficulty order relations (SQUARE), indirect difficulty order relations (IMPLY), implied item responses (MULT), and implied person-order relations (COUNT).

"Significance Tests"

The decision rule used in comparing frequencies is a rather liberal one. It has two parts. The major part corresponds to comparing frequencies by a binomial probability (McNemar's test; McNemar, 1969) and rejecting the null with a one-tailed alpha level of .33 . Values of $n_{jk}$ and $n_{kj}$ of 2 to 0 and 3 to 1, respectively, thus lead to rejection, and an implication is made.

The second aspect of the rule has to do with instances where the frequencies are 1 and 0. If the information is very sparse, i.e., early in the testing, most of the frequencies compared are either 0,0 or 1,0, and it is necessary to get some information out of the latter in order to improve item-assignments. However, when the information is less sparse, frequencies of at least 1,0 are almost bound to occur. Thus the decision as to whether to "believe" that a 1,0 frequency implies, say, a difficulty order, depends on how sparse the information is.

The specific nature of the test involves the evaluation of a chance probability. Suppose a vector has n elements, $n_j$ of which are 1 and the remainder zero. Suppose a second n-vector has $n_k$ ones and the rest zero. If the $n_j$ ones are scattered at random in the first vector and the $n_k$ are scattered at random in the other, and the two vectors are laid side by side, what is the probability that none of the ones in one vector are matched by ones at corresponding places in the other? If the complement of this probability is found, this is the probability of at least one pair of ones with the same index in the two vectors, i.e., a frequency that is not 0,0 . Obviously, if $n_j + n_k$ is greater than n, there must be at least one match. If not, the probability of zero matches

is given by the following formula, where $n_j$ is greater than $n_k$ :

$$p(0) = \frac{\binom{n - n_j}{n_k}}{\binom{n}{n_k}}$$

If p(0) is .5 or greater, this implies that a random match is unlikely to occur, and so the observed 1.0 frequency probably represents real information, and the implied relation is made. If it is less than .5, the probability is considered too great, that the matching elements occurred by chance, and no implication is made.

These standards will clearly seem incautious to anyone raised in the .05 - .01 tradition of significance testing. Two things may be borne in mind. Most of these implications of order are subject to reversal on the basis of later evidence, so the decisions are not irrevocable. Second, we do not have the same payoff matrix here as underlies traditional hypothesis testing. Particularly at the early stages, the penalty for concluding that there is not a difference in difficulty when there is one is as large as the penalty for concluding that there is a difference when there is not. We were, in fact, forced to abandon the use of more traditional significance levels by a good deal of exploratory simulation work, and it was not until we adopted this mode that we began to get reasonably good results.

## Assigning Items to Persons

At any given time in the testing process, as many inferences as possible are made about the relative difficulty of the items. These in turn

are used to imply responses for each person to items he has not yet taken, and these in turn are used to help determine the joint order of persons and items. The latter is necessary for the purpose of assigning items to persons optimally, insofar as current information allows, during the the course of the interactive testing.

A total score is assigned to each item and each person. For a person, this is the total number of items which he gets right, either directly or by implication (from MULT) plus the number of persons he ranks above or dominates (from COUNT) minus the number of items and persons that dominate him, as derived from the same sources. For an item, this total score is the number of persons who get it wrong--directly or indirectly as determined from MULT--plus the number of items it is harder than, determined from SQUARE and IMPLY, minus the number of persons who get it right, and the number of items it is easier than, as determined from the same sources. In this way, the persons and the items are ranked on the same scale.

The person takes the item for which he has no direct or implied response and which lies closest to him on the joint scale.

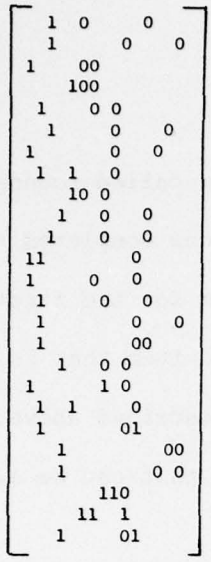
## Implementation

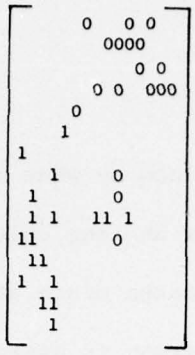
There are two basic modes of operation, which might be called simultaneous and cumulative. The simultaneous one, TAILOR, which was developed first (Cudeck, Cliff, Reynolds and McCormick, Note 1; Cudeck, Cliff, and Kehoe, in Press) assumes that a group of people is taking the test at the same time, whereas the cumulative one, TAILOR-APL, assumes that subjects are tested individually.

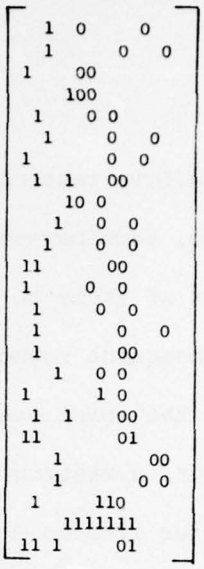
In TAILOR, testing takes place by what might be called rounds. At each round, each person receives an item unless he has completed the test. Assignment of items to persons takes place at random for the first round, and in subsequent rounds each person is assigned the item that is closest to him in the joint scale whose determination was described above. There is, however, a restriction on the number of persons that can be assigned a particular item on a particular round.

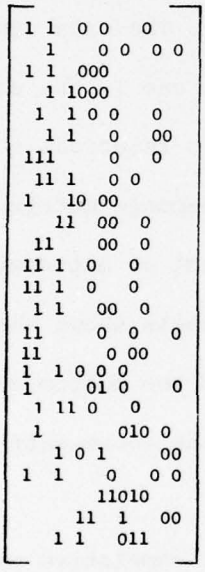
This procedure is illustrated in Figure 1 where the first three panels show the operation at an early stage of the process. The data are for 25 persons and 15 items on the Stanford-Binet. The left one is the actual response matrix; the middle one is the item dominance relations that are implied by them, and the right one is the implied response matrix. In each, a 1 means correct or dominance, a zero means incorrect or antidominance, and a blank means no relation. The middle set of panels shows the operation at an intermediate stage of the testing process, and the bottom one shows the final stage, the right most panel showing that the score matrix is now complete by implication.

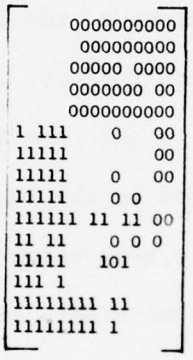
The second mode of operation is a sequential or cumulative one which tests individual subjects only. This is called TAILOR-APL (McCormick and Cliff, in press). Again, no knowledge about the items is assumed. The first person must therefore take all items. After a few persons, however, there may be enough information to define the relative difficulty of some items. Insofar as this is the case, it is used for subsequent persons to imply their responses to some items. As more and more information accumulates, more relative difficulty relations also accumulate, so that the tests become more and more "tailored" for later subjects.

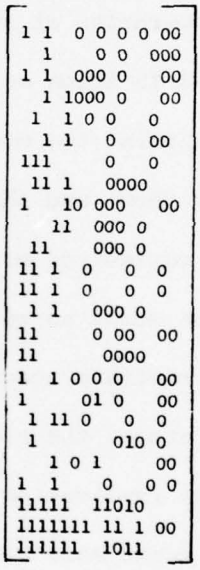


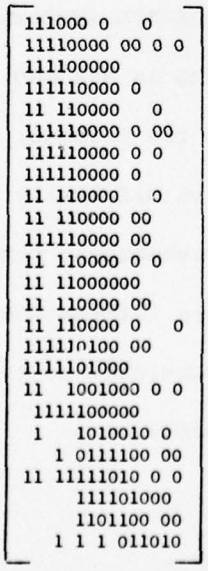


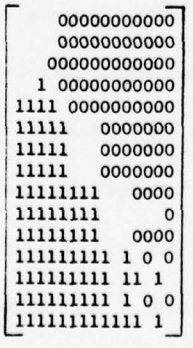












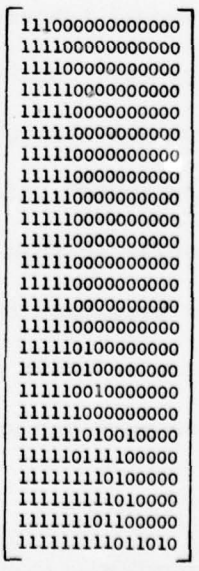


Figure 1. Response Matrix, item dominance matrix and implied response matrix for 3 rounds.

It is TAILOR which is evaluated in the two simulation studies reported here, while TAILOR-APL was evaluated with live subjects, and will be reported separately.

## Monte Carlo Study

### Two Types of Simulation

Considerable development and testing has gone into the development of these programs as workable operationalizations of the general concepts were sought. The TAILOR program seems to represent a satisfactory combination of characteristics, and a series of simulation studies were undertaken in order to assess its actual performance and its sensitivity to various parameters of the items and the testing situation.

Two types of simulation were done. The more extensive was carried out as a Monte Carlo process where the outcome of giving a particular item to a particular person was determined by calculating the probability of correct response according to the four-parameter Birnbaum model (Birnbaum, 1968) and comparing a random number on the 0,1 interval to that probability. The results of this study are the main focus of this report. Responses from human subjects were also used. A file of responses of 622 children, ages 2 years 0 months to 14 years 11 months, to the Stanford-Binet provided the data; correctness of a person's response to an assigned item was determined by looking in the file of responses.[1] The results of this study will be given in a later section.

### Parameters Studied in the Monte Carlo Study

In the simulation study, true score was defined to be normally distri-

[1]We wish to express our appreciation to Dr. Mark Reckase of the University of Missouri for providing this data file.

buted in the population with mean zero and standard deviation one. A sample of predetermined size was drawn from this population. Item difficulty was also assumed to be normally distributed, but the population mean and standard deviation could be varied as a program parameter.

This enabled us to study the effect of a mismatch between difficulty and true score. A predetermined number of items was sampled from the population with these characteristics. Thus, rather than being constrained to have the same mean and variance in the sample, item difficulty statistics could deviate as a result of either sampling fluctuations or deliberate manipulation of parameters. This was felt to introduce an element of realism that would be missing if difficulty and true score distribution parameters were constrained to be equal.

Discrimination values of the simulated items were also sampled from populations whose parameters could be varied. Again, a normal distribution was assumed for the discrimination parameters. In this way the effect of varying discrimination could be studied.

The guessing probability was also varied so that its effect could be assessed. It was fixed at a particular constant value rather than sampled from a population, however. The number of items in the pool and the number of persons being tested was also varied.

With this number of variables it was not possible to vary all of them simultaneously or over a wide range of values. Instead, two or three values of each were selected on the basis of realism and practicality, and small factorial designs involving two or three of them were constructed. In this way, the main effects and certain first and second-order interactions could be studied, but higher order interactions could not. The combinations selected

were chosen on the basis of their expected importance, and it is hoped that the principal effects were thus identified.

The analysis of the effects of these variables were carried out as analyses of variance and analyses of covariance, true score with observed score in the complete-data case for the same data being used as the single covariate. This permitted the assessment of the degree to which there were effects over and above that of the basic consistency of the data. In addition to the numerous small analyses, a regression analysis with all the data combined into a single score matrix and the main effects as independent variables was carried out, both with the covariate and without it.

The combinations of parameter values used here are given in Table 2. The assumed number of subjects was 10, 25 or 40, with the great majority of the data coming from the latter two values. The number of items was assumed to be either 15 or 25. These rather small numbers of persons and items were chosen for reasons of economy. Mean discrimination was assumed at either 1.0 or 2.0 except in two cases where it was .5 . The standard deviation of discrimination was usually assumed to be zero, but also took on values of .2 or .4 . The mean population difficulty was usually assumed to be zero, equal to the mean ability, but was also set at 1.0 for two runs. Similarly, the standard deviation of difficulty was usually 1.0, but set at 2.0 for two runs. The chance parameter was usually zero, but also took on values of .1 or .2 . The basic design was a 2 x 2 x 2 with 25 or 40 persons, 15 or 25 items, and discrimination of 1.0 or 2.0 . The other parameter variations were usually made singly in combination with different levels of one or two of these main parameters. The specific combinations used can be seen in Table 2, where each row defines a set of conditions for a sample

Table 2

Characteristics of Samples of Score Matrices

Generated by Latent Trait Models

Sample Characteristics

| Condition Number | Persons | Items | μ Discrimination | σ Discrimination | μ Difficulty | σ Difficulty | Chance |
|---|---|---|---|---|---|---|---|
| 1 | 10 | 25 | 1.0 | 0 | 0 | 1.0 | 0 |
| 2 | 10 | 25 | 2.0 | 0 | 0 | 1.0 | 0 |
| 3 | 25 | 15 | .5 | 0 | 0 | 1.0 | 0 |
| 4 | 25 | 15 | .5 | 0 | 0 | 1.0 | 0 |
| 5 | 25 | 15 | 1.0 | 0 | 0 | 1.0 | 0 |
| 6 | 25 | 15 | 2.0 | 0 | 0 | 1.0 | 0 |
| 7 | 25 | 15 | 1.0 | 0 | 1.0 | 1.0 | 0 |
| 8 | 25 | 15 | 2.0 | 0 | 1.0 | 1.0 | 0 |
| 9 | 25 | 15 | 1.0 | 0 | 0 | 2.0 | 0 |
| 10 | 25 | 15 | 2.0 | 0 | 0 | 2.0 | 0 |
| 11 | 25 | 15 | 1.0 | .2 | 0 | 1.0 | 0 |
| 12 | 25 | 15 | 2.0 | .2 | 0 | 1.0 | 0 |
| 13 | 25 | 25 | 1.0 | 0 | 0 | 1.0 | 0 |
| 14 | 25 | 25 | 2.0 | 0 | 0 | 1.0 | 0 |
| 15 | 25 | 25 | 1.0 | 0 | 0 | 1.0 | .1 |
| 16 | 25 | 25 | 2.0 | 0 | 0 | 1.0 | .1 |
| 17 | 25 | 25 | 1.0 | 0 | 0 | 1.0 | .2 |
| 18 | 25 | 25 | 2.0 | 0 | 0 | 1.0 | .2 |
| 19 | 25 | 25 | 1.0 | .2 | 0 | 1.0 | 0 |
| 20 | 25 | 25 | 2.0 | .2 | 0 | 1.0 | 0 |
| 21 | 40 | 15 | 1.0 | 0 | 0 | 1.0 | 0 |
| 22 | 40 | 15 | 2.0 | 0 | 0 | 1.0 | 0 |
| 23 | 40 | 15 | 1.0 | 0 | 0 | 1.0 | .2 |
| 24 | 40 | 15 | 2.0 | 0 | 0 | 1.0 | .2 |
| 25 | 40 | 25 | 1.0 | 0 | 0 | 1.0 | 0 |
| 26 | 40 | 25 | 2.0 | 0 | 0 | 1.0 | 0 |
| 27 | 25 | 25 | 2.0 | .4 | 0 | 1.0 | 0 |
| 28 | 25 | 15 | 2.0 | .4 | 0 | 1.0 | 0 |

of score matrices generated by the Birnbaum model. The correlation matrix of major variables appears in the appendix, Table A.

Data Generation

Five sample score matrices were generated according to the parameters of each line of Table 2. Given the person scores and the item parameters, a probability of correct responses of each persons to each item is computable using the Birnbaum model. Then a random number determined whether the response was correct or not. The resulting score matrices were stored for later reference by TAILOR. The specific procedures are in Appendix 1.

The intial round of assigning each person an item took place at random. The correctness of the response was determined by simply looking in the appropriate location of the stored matrix of model-generated item responses. For each subsequent round, the matching of item with person took place by means of the computation of implied responses and matching of item-person scores, as was outlined earlier. The session was complete when there was an actual or implied response for each item and person in the score matrix.

In addition to the true score that was used to generate the data, there are two scores for each person in a given sample. One is his score from the tailored simulation, and the other is his score on the complete test. These will be referred to as Tailored scores and Complete scores, respectively; Note that they are not experimentally independent because the responses on which the Tailored score is determined are a subset of the responses determining Total score. A number of statistics were calculated from these, and some of them were important to the later analysis.

Dependent Variables

A variety of independent variables were used, but not all in all analyses.

The correlations among True score, Tailored score and Total score were computed. The two correlations with True score are the validities of Tailored and Total scores. These correlations were used as a major dependent variable and a covariate, respectively. The reasoning was that agreement with true scores is the major quality which one desires from a tailored testing procedure and the expectation that variations in Total score validity would be the major source of influences which would need to be controlled. Both Pearson and rank-order (tau) coefficients were computed, but the results of the analyses were in extremely close agreement, so only one set will be reported, the tau. The Fisher Z-transform was not used because in our experience it has little effect on results unless the mean and variance of the correlations are both large, and rarely even then.

An additional variable reflecting the accuracy of the operation of TAILOR is the proportion of responses to items that were not taken which were correctly predicted by TAILOR. Other variables which were used reflected efficiency and cost of the procedure. In particular, the ratio of actual responses to total possible responses in a given score matrix is clearly relevant, as is the amount of computer processing time used per run. The effects of the independent variables on these were assessed as well as determining the overall levels.

A Monte Carlo evaluation of this kind with multiple dependent and independent variables provides opportunity for many analyses. Not all possible ones were performed. Rather, the attempt was to be selective and investigate the most important questions and assess the plausibility of the most reasonable alternative explanations for any effects that were present.

# Results

## Basic Results

While the major results are included in Table 3, particular portions of them will be selected out for further analysis and comment. The data there is listed in the same order of conditions as in Table 2. Here, the correlations, both Pearson and tau, of total score on the complete test with true score are given. Also given are the correlations with score from the TAILOR operation on the same data. The percentage of responses under TAILOR and the amount of CPU time used are likewise presented. The last line of the table shows the means of each variable.

The means represent averages across conditions which affect these variables, and some of the effects are substantial. Therefore, their precise values should be viewed rather tentatively. They do, however, provide a quick summary of the major findings. On the average, TAILOR presented 55 per cent of the items to each person. The validity correlations for the tailored scores averaged .757 (tau) and .889 (r). These are similar to the complete-test validities of .810 and .926.

## Influences on Proportion of Responses

Fifty-five percent of the items is rather a large proportion, if one is hoping to make material increases in test efficiency. Thus the factors that seem to influence the proportion of items in the test bank which must be presented to the person are of interest. The major ones appear to be the number of items and persons. The more items in the pool and the persons being tested, the smaller the proportion of items presented to each person. These results are not surprising; the larger

Table 3

Cell Means of Major Dependent Variables

| Condition Number | Complete Data Tau | Complete Data Pearson | Tailored Data Tau | Tailored Data Pearson | % Responses | CPU (per person) |
|---|---|---|---|---|---|---|
| 1 | .769 | .941 | .724 | .850 | .590 | 3.64 |
| 2 | .880 | .955 | .846 | .946 | .577 | 3.54 |
| 3 | .745 | .876 | .636 | .810 | .615 | .54 |
| 4 | .621 | .803 | .480 | .686 | .556 | .58 |
| 5 | .803 | .920 | .750 | .896 | .609 | 1.34 |
| 6 | .860 | .954 | .818 | .934 | .580 | 1.14 |
| 7 | .771 | .916 | .734 | .898 | .599 | 1.38 |
| 8 | .863 | .932 | .822 | .922 | .577 | 1.22 |
| 9 | .662 | .869 | .600 | .810 | .552 | 1.11 |
| 10 | .865 | .958 | .854 | .952 | .571 | 1.16 |
| 11 | .771 | .924 | .728 | .880 | .603 | 1.32 |
| 12 | .863 | .938 | .836 | .928 | .592 | 1.36 |
| 13 | .809 | .924 | .738 | .878 | .514 | 3.62 |
| 14 | .868 | .965 | .884 | .946 | .516 | 3.59 |
| 15 | .826 | .947 | .754 | .904 | .538 | 3.99 |
| 16 | .836 | .946 | .824 | .920 | .506 | 3.82 |
| 17 | .788 | .918 | .696 | .866 | .555 | 4.70 |
| 18 | .821 | .913 | .760 | .896 | .502 | 3.84 |
| 19 | .821 | .945 | .758 | .884 | .506 | 3.38 |
| 20 | .890 | .968 | .864 | .954 | .490 | 3.53 |
| 21 | .766 | .904 | .710 | .868 | .543 | 1.44 |
| 22 | .855 | .932 | .836 | .924 | .548 | 1.41 |
| 23 | .682 | .843 | .560 | .734 | .567 | 1.57 |
| 24 | .773 | .907 | .706 | .870 | .571 | 1.54 |
| 25 | .842 | .956 | .772 | .920 | .471 | 4.06 |
| 26 | .892 | .956 | .852 | .940 | .441 | 3.14 |
| 27 | .868 | .962 | .846 | .948 | .492 | 3.60 |
| 28 | .836 | .940 | .814 | .924 | .585 | 1.23 |

the item pool, the better a portion can stand for the whole. The strength of the effect of the number of persons is perhaps surprising, but it may be remembered that persons and items are treated symmetrically in the present theory. Another way to look at this is that more persons enable the program to get clear information on item difficulties more quickly. There is perhaps also a tendency for fewer responses to be required when the items are more discriminating.

These results are shown more clearly in Table 4. The upper part shows the mean proportion of responses for the 2 x 2 x 2 subset of the study which was mentioned earlier. Here, the strong effect of Items and Persons is quite apparent, but the effect of Discrimination is weak or non-existent. For the 25-item, 40-person cell, the mean proportion of responses is .456 . The lower section of the table shows the results of a regression analysis of all 28 conditions (n = 140) of proportion of responses on all three variables treated as main effects. There, the F column shows the significance of the regression weight for the individual variables, and Items and Persons are clearly signifcant but Discrimination is not.

The values used for these variables are limited by the practical considerations present in this study, and it would be interesting to explore a wider range of values. In particular, there must be diminishing returns in the effect of Persons and Items, but these data do not permit the assessment of the rate at which that takes place. Also, there should be a fairly strong effect of Discrimination if it reaches really low values since this would lead to a greater frequency of contradictory dominance relations with concomitant failure to pass the "significance" levels.

Table 4

Influences on Proportion of Items

Proportion of Items Used

in 2 x 2 x 2 Data Subset

| | 10 Persons | 25 Persons | | | 40 Persons | | |
|---|---|---|---|---|---|---|---|
| | Items | Items | | | Items | | |
| | 25 | 15 | 25 | Mean | 15 | 25 | Mean |
| Discr. = 1 | .590 | .609 | .514 | .562 | .543 | .471 | .507 |
| Discr. = 2 | .577 | .580 | .516 | .548 | .548 | .441 | .494 |
| Mean | .584 | .594 | .515 | .555 | .546 | .456 | .501 |
| | n = 2 | n = 20 | | | n = 6 | | |

Regression of Proportion of Items on

Persons, Items, and Discrimination

All Data

| Variable | b | beta | F | $R^2$ |
|---|---|---|---|---|
| Items | -.0076 | -.690 | 128.05 | .327 |
| Persons | -.0036 | -.459 | 56.69 | .524 |
| Average Discrimination | -.0101 | -.092 | 2.38 | .532 |

## Validity of Tailored Scores

The success of a tailored testing scheme is primarily measured by the ability of the tailored scores to substitute for scores on the complete test. In a Monte Carlo study like this where a true score is available, correlation with true score would appear to be the most appropriate criterion by which to judge the success of TAILOR.

As noted earlier, the Tailored validities are close to, but lower than, the complete data validities. The validities are also apparently, from Table 3, quite variable, ostensibly as a function of various parameters of the situation. A number of analyses were made to attempt to identify the characteristics which affect Tailored validity.

The primary effect is from the consistency of the actual sample of data. This is illustrated graphically in Figures 2 and 3 which plot mean Tailored correlation as a function of mean complete correlation for each of the 28 conditions. Figure 2, shows tau and Figure 3, r.

Treating each of the five replications under each condition separately, the correlation of complete and Tailored validities are .86 and .83 for tau and r, respectively. The slopes of the regression lines in the Figures are greater than unity, showing that a given effect on the validity of a complete test will have an even greater effect on the validity of the items in Tailored format.

## Influences of Independent Variables on Validity

A variety of regression analyses and analyses of variance and covariance were performed to identify influences on validity. Tailored Tau, the rank-order correlation of Tailored score with True score was

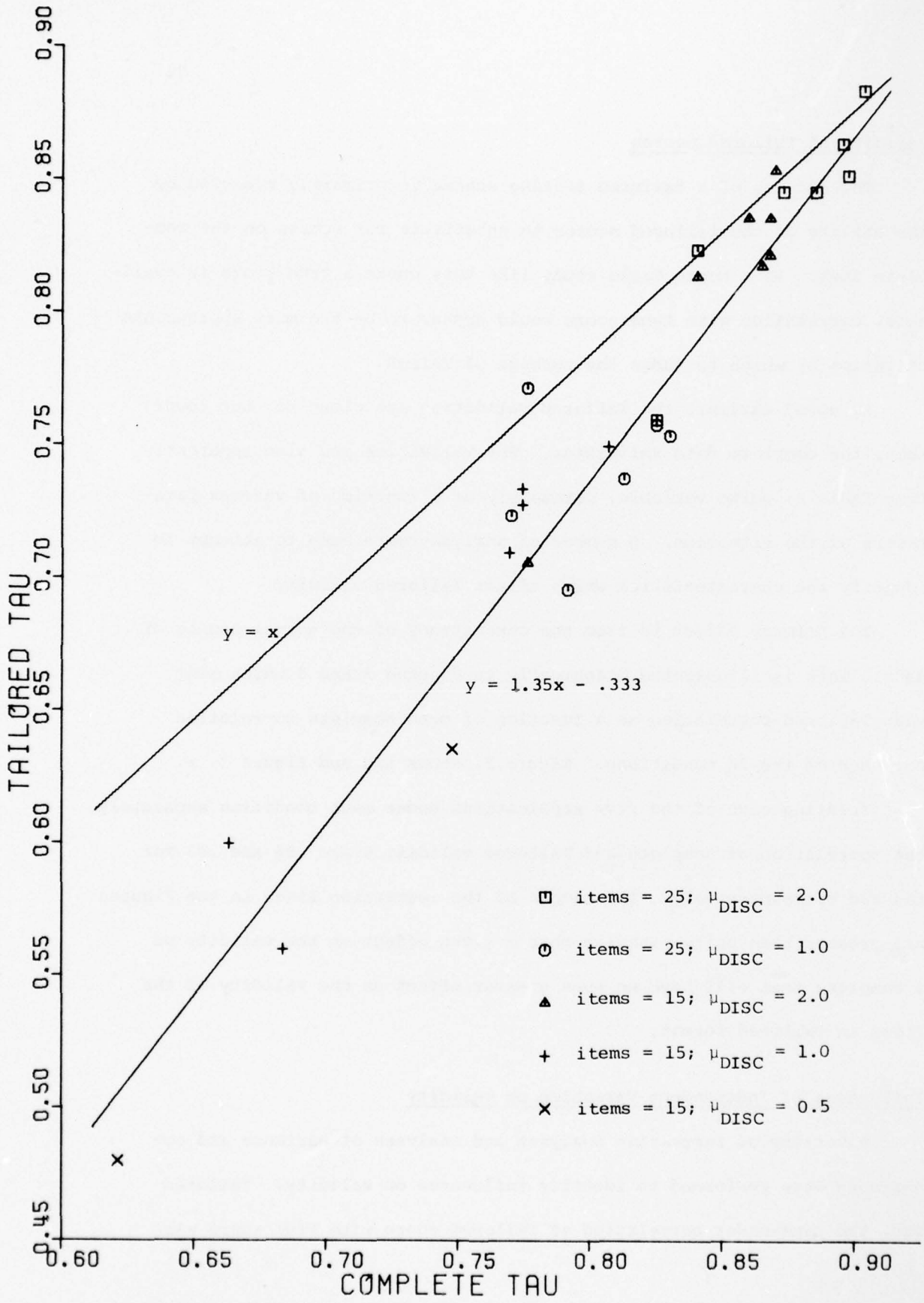



Figure 2. Complete test tau and tailored test tau for 28 studies using Birnbaum model

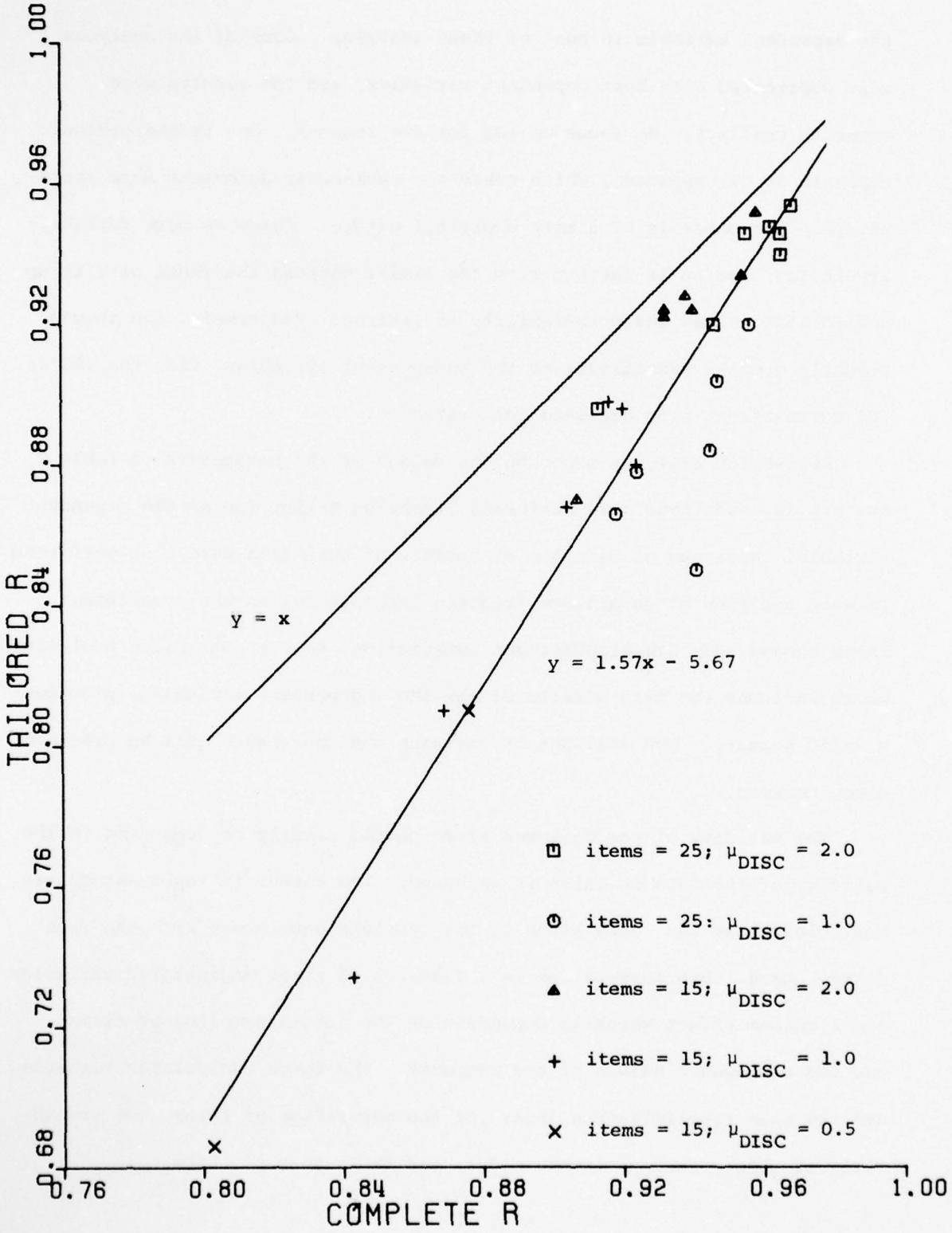



Figure 3. Complete test r and tailored test r for 28 studies using Birnbaum model

the dependent variable in most of these analyses. Some of the analyses were duplicated with both dependent variables, and the results were never in conflict. We focus on tau for two reasons. One is the ordinal emphasis of our approach, which makes the rank-order agreement more appropriate. The other is of a more empirical nature. There is more variability in tau, and it is farther from the limit, whereas the range of r is up toward high values where variability is limited. Futhermore, tau showed slightly greater sensitivity to the independent variables, viz. the .86 vs. .83 correlations with Complete Test validity.

Regression analyses based on the values of the parameters in Table 2 for all 28 conditions were performed, treating Tailor Tau as the dependent variable. Analyses of variance of subsets of that data were also performed as were analyses of covariance treating Complete Tau as the covariate. These showed only one significcant interaction, so the regression analysis, which includes the main effects of all the independent variables, provides a valid summary. The analyses of variance and covariance will be presented also, however.

The validity of the Tailored score should clearly be dependent on the validity of the data on which it is based. The latter is represented here most clearly as the correlation of the Complete data score with the True score. Here, this correlation is a function of three manipulated variables and a random effect which is dependent on the actual sampling of items and the stochastic nature of the response. The three manipulated variables are the mean discrimination index for the population of items, the probability of chance success in the model, and the number of items.

Table 5 shows the results of stepwise regression analyses of Tailored Tau on predictors. The analyses were performed using the SPSS package (Nie, Null, Jenkins, Steinbrenner, and Bent, 1975). The upper section shows the reults when Complete Tau is included, and the lower when it is not. The "Significant Variables" are those which significantly increased the multiple R as they were included. After they were included, none of the remaining ones were shown as adding significantly to the multiple R. (There is a possibility with such a procedure that some combination of the not-included variables might have added singificantly, but this possibility is remote and often unstable when it occurs.)

Complete Tau was the first variable entered in the regression equation, accounting for 74 per cent of the variance of all 140 observations. However, two of its causes, Mean Discrimination and Chance Probability, added significantly to the multiple, contributing five and two per cent of variance, respectively. All three have weights of the expected sign: higher Complete Tau and higher Mean Discrimination lead to higher Tailored Tau, and greater Chance Probability leads to lower.

When only the manipulated variables are included, the third influence on Complete test validity also becomes significant. Now, Mean Discrimination accounts for almost half the variance (49%); Chance Probability contributes another eight, and the number of items a final five percent. The total percentage of variance accounted for is 62 rather than 82 as it was when the consistency of the actual data set was included. The number of persons, the variability in discrimination of the items, and the mean standard deviation in difficulty are not significant influences on validity with the levels of parameters used here. The level of pre-

Table 5

Regressions of Tailored Tau

on Independent Variables

| Variables | b | beta | F | $R^2$ |
|---|---|---|---|---|
| I. Including Complete Tau | | | | |
| Significant Variables: | | | | |
| Complete Tau | .8124 | .6400 | 183.64 | .743 |
| Mean Discrimination | .0585 | .3270 | 49.32 | .796 |
| Chance Probability | -.2366 | -.1525 | 16.15 | .818 |
| Nonsignificant Variables: | | | | |
| Items | | | | |
| Persons | | | | |
| Standard Deviation of Discrimination | | | | |
| Standard Deviation of Difficulty | | | | |
| Mean Difficulty | | | | |
| II. Manipulated Variables Only | | | | |
| Significant Variables: | | | | |
| Mean Discrimination | .1215 | .6794 | 162.94 | .490 |
| Chance Probability | -.4910 | -.3166 | 35.68 | .572 |
| Items | .0053 | .2362 | 19.39 | .625 |
| NonSignificant Variables: | | | | |
| Persons | | | | |
| Standard Deviation of Discrimination | | | | |
| Standard Deviation of Difficulty | | | | |
| Mean Difficulty | | | | |

cision here (witness the small proportion of variance for Items) is such that these variables must only have small effects, if any. The lack of negative effects for numbers of persons and items is particularly interesting in view of the fact that a smaller proportion of responses are used as these increase.

The 28 conditions contain a number of subdivisions which represent orthogonal designs representing two or three of the manipulated variables. These allowed the investigation of a number of first-order and a few second-order interactions among the variables as well as the highlighting of the main effects. That is, there are a number of small factorial designs contained in the total set; the other variables are held constant at some particular combination of levels.

One such subdesign is shown in Table 6. The analysis of variance (anova) and the analysis of covariance (ancova) show that there is an effect for interaction. Even when corrected for the Complete Tau as a covariate, changing the discrimination index from 1.0 to 2.0 will increase the Tailored Tau by about .06. Not considering the matrix consistency, i. e., the actual effect of changed discrimination in practice would increase validity by about .10.

Two combinations of conditions cross Mean Discrimination with Chance Probability. One varied Chance Probability at the levels 0, .1, .2 and Mean Discrimination at 1.0 or 2.0 with 25 Items and 25 Persons. The other used only the two extreme levels of Chance (0 and .2), used the same levels of Mean Discrimination, and held Persons and Items constant at 40 and 15 respectively. Means for the conditions are shown in Table 7.

Both sets of Table 7 data show main effects for both variables and non-

Table 6

Effect of Persons, Items

and Discrimination on Validity[a]

| Mean Discrimination = 1 | | | Mean Discrimination = 2 | | |
|---|---|---|---|---|---|
| | Persons | | | Persons | |
| Items | 25 | 40 | Items | 25 | 40 |
| 15 | .750 | .710 | 15 | .818 | .836 |
| 25 | .738 | .772 | 25 | .884 | .852 |

$F_p = .09$ ; $F_I = 3.74$ ; $F_D = 37.87$: All interactions $> .05$

[a] S. D. Discrimination = 0; S. D. Difficulty = 1.0; Chance Probability = 0; Mean Difficulty = 0.

Table 7

Effect of Discrimination and Chance

Probability on Validity[a]

| Persons = 25, Items = 25 | | | | Persons = 40, Items = 15 | | |
|---|---|---|---|---|---|---|
| | Chance Probability | | | | Chance Probability | |
| Discrim. | 0.0 | 0.1 | 0.2 | Discrim. | 0.0 | 0.2 |
| 1.0 | .738 | .754 | .696 | 1.0 | .710 | .560 |
| 2.0 | .884 | .824 | .760 | 2.0 | .836 | .706 |
| $F_D = 14.19$; $F_C = 4.01$ | | | | $F_D = 17.69$; $F_C = 18.75$ | | |

[a] S. D. Discrimination = 0; S. D. Difficulty = 1.0.

significant interaction, by analysis of variance. In the 2 x 3 case, Mean Discrimination is significant at .01 and Chance at .05. In the other, both are significant at .01 . Slight differences in the relative sizes of the effects in the two cases lead to different significant effects when the covariate is used. In the 2 x 3 case, Mean Discrimination is significant at .01 while Chance is not significant (although F = 1.94). In the 2 x 2 data, the situation is reversed. Chance Probability is significant at .05, whereas Mean Discrimination is not (although F = 3.59). It seeems likely that both have a real effect (Cf. the regression analysis) but that random fluctuations of small magnitude led to the failure of one to reach signifi-cnance in one case while the other failed in the second. The use of only the extreme levels of Chance Probability in the second case may have contributed. The alternative is that there is a high-level interaction with Items or Persons or both. Unfortunately, the choice levels confounds the persons and items in these data. Full explication would require using a four-variable factorial, which did not seem justified. If these uncertainties are set aside, it appears that items with an inherent discrimination parameter of 1.0 and a zero guessing probability will behave about as well as items with a discrimination parameter of 2.0 and a guessing probability of .2, i. e., five-choice items.

The effect of discrimination indices when they vary over a wider range is shown in Table 8, which includes a .5 discrimination index as well as 1.0 and 2.0. Note that all other variables are fixed at certain levels, of which the most important is that Chance Probability is zero. Also, the lower levels of Persons and Items are used. Lowering discrimination from 1.0 to .5 has an even stronger effect here than the step from 2.0 to 1.0,

Table 8

Effect of Lowest Mean

Discrimination on Validity[a]

| Mean Discrimination[a] | | |
|---|---|---|
| .5 | 1.0 | 2.0 |
| .636 | .750 | .818 |

$F_D = 11.91$

[a]Persons = 25; Items = 15; Chance Probability = 0;
S. D. Difficulty = 1.0; Mean Difficulty = 0;
S. C. Discrimination = 0.

Table 9

Effect of Lowest Levels of Persons

on Validity[a]

| Mean Discrimination | Number of Persons | | | |
|---|---|---|---|---|
| | 10 | 25 | 40 | Mean |
| 1.0 | .724 | .738 | .772 | .745 |
| 2.0 | .846 | .884 | .852 | .861 |
| Mean | .785 | .811 | .812 | .803 |

$F_D = 26.62$ ; $F_P = .62$

[a] Items = 25; Chance Probability = 0; S. D. Difficulty = 1.0;
Mean Difficulty = 0; S. D. Discrimination = 0

apparently, although the latter difference happens to be smaller here than it is in most of the data. These differences remain, even increase slightly, in the analysis of covariance, although there the significant level just misses the .01 criterion ($R^2$ = .67) whereas, rather anomolously, this is reached in the analysis of variance.

As an experiment in the robustness of the system, the number of persons was reduced all the way to ten in the next set of data. As shown in Table 4, the proportion of items is somewhat increased, but is still below 60 per cent with 25 items. Table 9 shows that there is essentially no effect on validity, and the analyses of variance and covariance confirmed this; both Persons and Persons by Discrimination interaction were far from significance. There is presumably some effect on efficiency, because more items are asked, but this balances out to give equivalent validity for the final result. Discrimination has its usual strong effect here. The important thing is that the TAILOR system will apparently operate adequately with as few as ten persons, provided item discrimination reaches these acceptable levels and the guessing probability is negligible.

Monte Carlo studies typically assume a constant value for the discrimination index, whereas in actuality items are variable in their discriminatory power. The next set of data investigated the effect of variable discriminatory power (S. D. Discrimination) on the validity of TAILOR scores. Standard deviations of .2 and .4 were used, the latter only with Mean Discrimination of 2.0. Thus the manipulations were relatively mild since all items sampled would still have substantial positive discrimination. With these levels, there is no effect whatever, as can be seen in Table 10. The analyses of variance and covariance confirmed this. Because of the unbalanced nature

Table 10

Effect of Variation in Discrimination, Mean Discrimination, and Number of Items on Validity[a]

| | Mean Discrimination = 1.0 | | | | Mean Discrimination = 2.0 | | | |
|---|---|---|---|---|---|---|---|---|
| | S. D. Discrimination | | | | S. D. Discrimination | | | |
| Items | 0 | .2 | .4 | Mean | 0 | .2 | .4 | Mean |
| 15 | .750 | .728 | -- | .739 | .818 | .836 | .814 | .827 |
| 25 | .738 | .758 | -- | .748 | .884 | .864 | .846 | .865 |
| Mean | .744 | .743 | | .744 | .851 | .850 | .830 | .844 |

[a]Persons = 25, S. D. Difficulty = 1.0; Chance Probability = 0; Mean Difficulty = 0

of the design, two analyses were done, one of the 2 x 3 with Mean Discrimination of 2.0, on the right, and one of the 2 x 2 x 2 which is left when S. D. Discrimination of .4 is deleted. All show tiny mean squares for S. D. Discrimination. The usual effect of Mean Discrimination is present, and there is a .05 level main effect for items in the anova which washes out in the ancova. No interactions approach significance. Thus, the TAILOR procedure is not sensitive to mild variations in the discriminating power of the items in the pool being used.

The final set of analyses attempted to assess the effect of mismatches of the item difficulty for the population tested. In almost all the data used here, the item difficulty was normally distributed with a mean of 0 and variance 1.0, i.e., difficulty has the same distribution as true score. The actual items simulated were sampled from these populations, so they would have characteristics that deviated from the population values due to sampling. These last sets of data dealt with items which deliberately deviated. In particular, items from a population with a mean difficulty of .5 (and a variance of 1.0) were sampled. These data are in the left section of Table 11. Additionally, a second set with a difficulty standard deviation of 2.0 were sampled.

The moderate deviation in mean difficulty had no significant effect in either the anova or the ancova. On the other hand, Table 11 shows appreciable differences for S. D. Difficulty, primarily as an interaction with Mean Discrimination. The main effect is not significant ($F = 2.86$), but the interaction is significant at .01, indicating that items which are highly variable in difficulty work better than moderately variable ones when discrimination is high, but the reverse is true when discrimin-

Table 11

Effects of Difficulty Parameters on Validity[a]

| Mean Discrimination | Mean Difficulty 0 | .5 | Mean Discrimination | S. D. Difficulty 1.0 | 2.0 |
|---|---|---|---|---|---|
| 1.0 | .750 | .734 | 1.0 | .750 | .600 |
| 2.0 | .818 | .822 | 2.0 | .818 | .854 |
| | $F_{DISC} = 7.03$ ; $F_{DIFF} = .04$ | | | $F_{DISC} = 22.84$ ; $F_{DIFF} = 2.86$ | |

[a]Items = 15; Persons = 25; Chance Probability = 0; S. D. Discrimination = 0.

Table 12

Central Processing Time in Seconds by Items and Persons

| | Persons | |
|---|---|---|
| Items | 25 | 40 |
| 15 | 31.3 | 59.5 |
| 25 | 94.7 | 143.9 |

ation is merely good. This effect disappears in the ancova. Apparently it is virtually entirely attributable to the basic validity of the data generated under the various conditions. In fact, in this particular case there is no effect for Mean Discrimination in the ancova.

Computer Time

A real-time system is only useful if it can operate in real time, and a computerized testing system cannot be too costly if it is to be adopted. The amount of central processing unit time (CPU) used in each computer run was recorded as part of the operation of the program. A few conditions that were anomalous for technical programming reasons were deleted, and the average CPU for each of the main combinations of Persons and Items were computed. These are given in Table 12 where it is apparent that both have a substantial effect, particularly Items. Pro-rated across persons, this indicates that about four seconds of CPU is expended per subject with a pool of 25 items. This is admittedly on a highly efficient 370/158 installation, but at charges which are currently about five cents per CPU second, computing costs do not seem to be a major factor. It should be noted as well that a good part of the computer time went for overhead routines which were used to monitor the process and would not be included in an operational version of the program. Furthermore, we foresee substantial increases in program efficiency in the near future, and computing costs seem to be continuing their historic decline, rather than leveling off. Therefore, we do not foresee computing considerations being a major factor with item pools of substantially larger size.

Summary of Results of the Monte Carlo Study

It appears that under a variety of circumstances the TAILOR procedure works quite well. Without any pretesting, it arrives at a reasonable approximation to the total score on a test, using about half the items. Moreover, the percentage decreases with the number of persons and items without there being a concommitant loss in validity. The major determinant of the validity of the Tailored score is the validity of the item responses on which it is based. It is somewhat more sensitive to influences on consistency such as Mean Discrimination and Chance than the total score is. It is relatively robust with respect to variations in a variety of parameters, and computationally efficient enough for practical use, provided the items are of the levels of quality used here.

## Data Bank Simulation

### Data Source

Responses from human subjects were also used in simulation studies of TAILOR in order to see if the results generated from the Birnbaum model (Birnbaum, 1968) would carry over to human responses. The approach here was to make use of a data bank consisting of the complete item response matrix for a large sample of persons. Specifically, the data was a file of responses of 622 children to the 122 items of the Stanford-Binet, which was made available through the kind offices of Dr. Mark Reckase of the University of Missouri.

The children ranged in age from 24 months to 178 months with a mean of 93.7 and a standard deviation of 40.6 months, with a more or less uniform distribution. The mean IQ was 117.3 with a standard deviation of 17.6 and a range from 66 to 166. Thus, the sample was well above average but quite variable in IQ.

The total sample was first divided into four age ranges: 24 to 59, 60-95, 96-131, and 132-179. The three younger groups represent three-year spans and the oldest is a four-year one. This was done in order to reduce the variance in ability. The 96 to 131 age group was felt to be unlikely to provide additional information and was not used in any of the simulations.

Within each age group, item statistics were calculated, i.e., proportion correct and the discrimination parameter (Urry, 1974). Item pools were formed for each age group on the basis of the item difficulty by deleting any subtest whose items were <u>all</u> of 1.0 or 0.0 difficulty. Of the 122 items on the total test, this left 54, 72, and 74 in the respective age groups. These characteristics of the age groups and item pools are given in Table 13.

Table 13

Characteristics of Age-Group

Binet Data

| Age (years) | N | Mean Discrimination | Binet Levels | Number of Items | Mean Difficulty | S. D. Difficulty |
|---|---|---|---|---|---|---|
| 2-4 | 156 | .70 | 2-6 to 8-0 | 54[a] | .095 | 1.84 |
| 5-7 | 179 | 1.71 | 4-0 to 14-0 | 72[b] | -.287 | 1.84 |
| 11-14 | 130 | 1.26 | 7-0 to Ad.3 | 74[a] | .562 | 1.59 |
| 3 | 64 | | 2-6 to 8-0 | 54 | | |
| 6 | 62 | | 4-0 to 14-0 | 72 | | |
| 13-14 | 60 | | 7-0 to Ad.3 | 74 | | |

[a] Includes 2 items with zero variance, not used to calculate mean, S. D.

[b] Includes 1 item with zero variance, not used to calculate mean, S. D.

The "testing" situation thus at last approximates to some degree one where children of a broad but limited age range are given a test whose items are of a broad difficulty range but are not completely inappropriate. The mean discrimination indices are of interest. For the two older groups, their values seem to approximate the 1.0 to 2.0 values that were the main ones used in the Monte Carlo studies, while for the youngest group they average below 1.0. The main difference from the Monte Carlo is perhaps that there is more variance in item difficulty here.

In addition to these three groups, three others that were more homogeneous with respect to age were used. One of these was all the three-year olds, one was the six-year-olds, and one was the thirteen and fourteen-year-olds. These too are listed in Table 13, but it should be noted that the information on items was not separately derived for them (samples were too small). Curtailing the range in age will further reduce ability variance, so in effect the discrimination of the items is reduced over the values in the larger, more variable groups.

The major purpose here is to observe the operation of TAILOR with data from human subjects. The operating characteristics such as average proportion of items used are of interest, but it would be surprising if they were very different from those for comparable numbers of subjects and items from the Monte Carlo study described earlier. The primary interest is in the effectiveness of the Tailored score. This was measured by its validity against true score in the Monte Carlo, but here there is no true score. In its absence, a parallel form reliability is the most logical index of quality, so it will be the variable of major interest.

## Method

Each age group was treated separately, so the basic data source was the matrix of item responses appropriate to it. For example, in the 2-4 group, this was the responses of the 156 members of the group to the 54 items that were in the appropriate subtests of the Stanford-Binet. Sample sizes of 20 and 40 were used, always with 25 items. There were five replications for each age-group and each sample size. In each replication a random sample of persons was selected as were two non-overlapping samples of 25 items. One sample of items in a replication was used only for the calculation of a total score. The other was used as a source of item responses for TAILOR, which operated in the same way as in the Monte Carlo investigations, using these as the response matrices. This procedure gave a Tailored and a Complete score for each person, each based on non-overlapping random samples of items from the same pool. Pearson and tau correlations were computed between Complete and Tailored scores and the correlations were averaged over the five replications. In addition, average correlations between Complete test scores on random subsets of 25 items were computed so that the Complete-Tailor correlations could be compared to them. Thus there were two parallel-form reliabilities, a Complete-Tailor and a Complete-Complete.

## Results

TAILOR behaved very similarly to the Monte Carlo, as far as apparent mode of operation of the program and number of responses required were concerned. Table 14 gives the average proportion of responses required of each subject for each of the conditions. These are very similar to those for the most similar conditions in the Monte Carlo, perhaps very slight-

Table 14

Proportion of Items

Used by TAILOR

| | Wide Age Range | | | | | Narrow Age Range | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Age - Range (Years) | | | | | Age - Range Years) | | | |
| | 2-4 | 5-7 | 11-14 | Mean | | 3 | 6 | 13-14 | Mean |
| n = 20 | .540 | .560 | .556 | .552 | n = 20 | .540 | .560 | .564 | .555 |
| n = 40 | .444 | .452 | .468 | .455 | n = 40 | .464 | .488 | .488 | .480 |

Table 15

Parallel Form Correlation (tau)

Complete and Tailored Tests

I. Wide Age Range

| | 20 Persons | | | | | 40 Persons | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Age - Range (Years) | | | | | Age - Range (Years) | | | |
| | 2-4 | 5-7 | 11-14 | Mean | | 2-4 | 5-7 | 11-14 | Mean |
| Complete-Tailor | .692 | .735 | .695 | .707 | C-T | .726 | .708 | .724 | .719 |
| Complete-Complete | .717 | .741 | .739 | .732 | C-C | .760 | .759 | .765 | .761 |

II. Narrow Age Range

| | 20 Persons | | | | | 40 Persons | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Age - Range (Years) | | | | | Age - Range (Years) | | | |
| | 3 | 6 | 13-14 | Mean | | 3 | 6 | 13-14 | Mean |
| Complete-Tailor | .565 | .616 | .668 | .616 | C-T | .583 | .601 | .667 | .617 |
| Complete-Complete | .646 | .652 | .733 | .677 | C-C | .636 | .664 | .727 | .676 |

ly higher. The indication is that TAILOR completes the score matrix on the basis of responses to about half of the 25 items, slightly more if there are 20 persons, somewhat less if there are 40.

Table 15 gives the parallel form correlations for Complete-Tailor and Complete-Complete scores. In the wide age range groups, these are very similar taus averaging .713 vs. .747 overall. The difference is somewhat larger in the narrow age ranges, .616 vs. .676. There is no consistent effect for either Age Range or Persons in the Wide Age Range data, but there does seem to be one in the Narrow Age Range. However, overall it is clear that there is a close correspondence between Complete-Tailor (C-T) and Complete-Complete (C-C) reliabilities.

This is true in a correlational sense as well as in an overall correspondence of averages, as is displayed graphically in Figure 4 where C-T is plotted as a function of C-C for the 12 means of Table 15. There is clearly a high correlation (.95), and again a slope greater than unity (1.17). As in the Monte Carlo data, the major influence on Tailored reliability is the consistency of the basic data, and the influence is disproportionate; that is, a given change in complete test reliability will produce an even larger change in Tailored reliability, just as was true for validity in the Monte Carlo.

In the Monte Carlo data, it was found that the two manipulated factors, Mean Discrimination and Chance Probability, had an effect over and above the effect of the validity of the complete test. A similar effect is apparent here in that analysis of covariance shows significant intercept differences between the Wide and Narrow range data, although the differences are not large. Note, for example, that the two Narrow range points with C-C

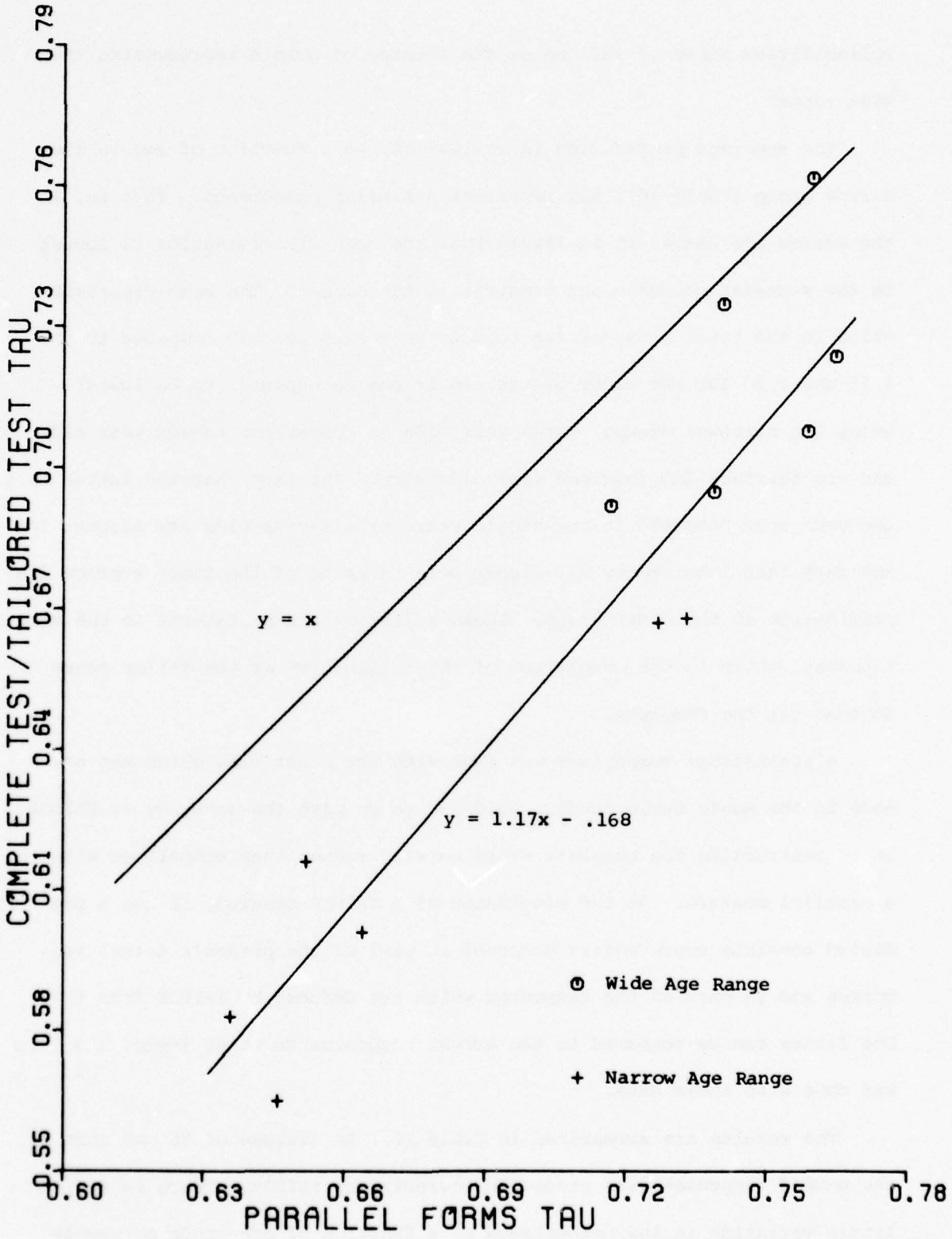



Figure 4. Complete test-tailored test reliability and Complete test-complete test reliability for 12 Binet studies

reliabilities above .7 fall below the cluster of points representing the Wide range.

The apparent progression in reliability as a function of age in the Narrow group (Table 16), may represent a similar phenomenon. That is, in the Narrow age range, it is likely that the mean discrimination is lowest in the youngest children and greatest in the oldest. The mean discrimination in the total group of two to four year olds was .70 compared to the 1.71 and 1.26 for the older groups, so it can be expected to be lowest among the narrower groups, three year olds in comparison to six year olds and the thirteen and fourteen year olds also. The fact that the latter is a two year span compared to the single year for six-year olds may account for the fact that consistency was higher here in spite of the lower average discrimination in the older group. These relations are peripheral to the main interest, which is the comparison of the reliability of the Tailor score to that for the Complete.

A statistical comparison was made with the Binet data which was not made in the Monte Carlo study. This had to do with the accuracy of TAILOR in reconstructing the complete score matrix, rather than comparison with a parallel measure. At the conclusion of a TAILOR session, it has a predicted complete score matrix composed in part of the person's actual responses and in part of the responses which are deduced by TAILOR from them. The latter can be compared to the actual responses to these items, and this was done with these data.

The results are summarized in Table 16. An average of 96 per cent of the unused responses were predicted correctly by TAILOR. There is very little variation in the percentages as a function of age-group or sample size. Since there are typically 10 to 15 items which an individual does

Table 16

Proportions of Responses

Correctly Predicted by TAILOR

| | Wide Age Range | | | | | Narrow Age Range | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Age - Range (Years) | | | | | Age - Range (Years) | | | |
| | 2-4 | 5-7 | 11-14 | Mean | | 3 | 6 | 13-14 | Mean |
| n = 20 | .950 | .977 | .960 | .962 | n = 20 | .933 | .977 | .940 | .950 |
| n = 40 | .957 | .956 | .951 | .955 | n = 40 | .951 | .969 | .953 | .958 |

not take, this means that in the majority of persons all of the responses are predicted correctly and in almost all of the remainder one or two are missed by TAILOR. These few errors are apparently sufficient to drop the correlation with a parallel form by a few points. It is clear, however, that TAILOR is quite accurate at predicting the outcomes of giving items.

Summary of Binet Simulation

The Binet data behaved very similarly to the Monte Carlo simulation. The item statistics indicated that the Stanford-Binet items were grossly similar to the values used in the Monte Carlo, and the proportions of responses used were also similar. Given the differences that the main dependent variable here was a parallel form correlation rather than a validity, the results of comparison of Complete to Tailored correlations were quite comparable in the two studies. Tailored reliabilities average within a few points of Complete ones, and show similar effects for the consistency of the basic data matrices. Moreover, TAILOR is highly accurate at reproducing the actual responses to items which it does not directly observe.

## Discussion

Reported herein are the results from two extensive studies using TAILOR with artificial data. In the first Monte Carlo investigation 3800 "persons" and 2750 "items" were generated with a mean ratio of validities against true scores for complete data and validities for tailored data equal to .933. Using the Binet data bank, 1800 "persons" taking 1500 "items" were simulated, and the ratio between alternate forms reliability with complete data ($r = .712$) and alternate forms with one complete test and a second tailored test

(r = .665) was very similarly .934 . This is an interesting finding, for each simulated person and item carried no information about its ability or difficulty. The usual requirement of pre-testing, which for stable estimates requires substantial samples, has been circumvented. More important, the accuracy of the test results was high. The implications of these findings in a tailored testing context are discussed below.

Efficiency of TAILOR

How much does a tailored test save? One answer to this can be found by comparing the reliability of a tailored test to one which is simply shortened to an equivalent length. Alternatively, given the reliability of tailored and complete tests, one can solve the Spearman-Brown formula (Lord and Novick, 1968, p. 112) for the length factor and compare it to the actual proportion of items asked in the tailored version.

The latter was done for the Monte Carlo data, starting by squaring the validities to get a reliability estimate. The Pearson correlations were used for this purpose. The most representative case is the data for 25 items and 40 persons, with discriminations of 1.0 and 2.0, conditions 25 and 26 of Table 3 where Tailored validities are .920 and .940, respectively, while the Complete validities are .956 and .965. The relation of .965 to .940 corresponds to using a test 78.4 per cent as long, whereas in actuality, only 44.1 per cent of the responses were used by TAILOR. A test that requires only an average of about 11 responses is acting like a complete test with nearly 20. The corresponding data with discriminations of 1.0 is not as favorable, but still is encouraging. Here, the validities of .956 and .920 correspond to using a test 72.6 as long, while in fact TAILOR used

47.1 per cent of the possible responses. Here, 12 responses are acting like an 18-item test. More exactly, the ratio of actual responses to lengths estimated from reliability is 1.778 in the case of the 2.0 discrimination item pool and 1.541 in the case of 1.0 discrimination.

The corresponding calculations for the smaller item pools are not as favorable. For 15 items, the ratios of number of responses to lengths estimated from reliabilities are 1.612 for 2.0 discrimination and 1.257 for 1.0, primarily becuase of the larger proportion of items that are used. The relation between the results for 25 and 15 items suggests that the savings for item pools of a more realistic size will be even more substantial. That is, the process becomes relatively more efficient as the number of items increases because the items that are used are on the average closer to the person's ability levels.

Parallel calculations were done for the Binet data. Here, though, the correlations are between a tailored test and a complete test, analogous to correlations between a long form and a short form, rather than between forms of equal length. This requires an adaptation of the Spearman-Brown formula which becomes:

$$r' = \frac{k^{\frac{1}{2}} r}{((k - 1)r + 1)^{\frac{1}{2}}} \quad ;$$

where r is the correlation between two parallel forms of equal lengths and r' si the expected correlation of one of them with a measure of the same ability which is of an altered length, k times the length of the first pair.

Given the correlations as here, the formula may be solved for k

$$k = \frac{r'^2 - r'^2 r}{r^2 - r'^2 r} .$$

This gives us a figure which can be compared to the proportion of items actually used, on the average, by TAILOR, as was done more simply with the Monte Carlo results.

The relevant data is given in Table 17, the proportion of items used, the Complete-Tailor and Complete-Complete correlations, the k values from the formula above, and the ratio of k to the proportion of items.

The results show that there is an appreciable gain in efficiency in the Wide age range, but only a small one in the Narrow. These results are similar to, but not as favorable as, those for the Monte Carlo. The efficiency ratio of 1.479 for TAILOR using 25 items and 40 persons is close to the 1.541 for 25 items of 1.0 discrimination derived from the Monte Carlo. There may be further increases in efficiency with a larger item pool, but on the other hand the Binet is an extremely discriminating test.

## Assessment Procedures

As is true of all areas of research, the assessment of tailored testing procedures is subject to bias and confounding. The attempt has been made here to make a realistic assessment of TAILOR. The Monte Carlo correlations with true score, based on results from a sample of items from a specified population parallel the use of TAILOR on a sample of persons with items which should be of appropriate average difficulty but may not be exactly so. The items' individual characteristics are unknown. TAILOR was compared to just using a random fraction of the items for all persons,

Table 17

Efficiency Calculations for

Binet Data

| | 20 Persons Wide | 40 Persons Wide | 20 Persons Narrow | 40 Persons Narrow |
|---|---|---|---|---|
| % Responses | .552 | .455 | .555 | .480 |
| C-T r | .829 | .866 | .765 | .750 |
| C-C r | .855 | .889 | .811 | .805 |
| k | .695 | .673 | .604 | .562 |
| ratio k/% | 1.259 | 1.479 | 1.088 | 1.171 |

and found to be appreciably more efficient, particularly with highly discriminating items.

Note that questions that can be raised about tailoring procedures based on pretest item statistics, questions concerning the sampling errors in the statistics and possilbe population differences between pretesting and tailored populations, are not relevant here. Also, there is no hidden cost of pretesting. After all, if a test must be standardized on 1,000 subjects so that a second thousand need only take half the items, the average person has actually taken three-fourths of the items, not half.

In the Binet data, where there is no true score, the correlational design furnished information on the efficiency of TAILOR using parallel measures. The comparison to the complete test data is confounded with differences in the item pools for tailored and conventional tests. The estimates of the accuracy with which the partial response patterns on TAILOR can reproduce the complete score matrix are not biased by the inclusion of the responses themselves.

## Theoretical Considerations

The ordinal model that is the basis for TAILOR is a rather different appraoch to test theory. One of its advantages is that it treats items and persons symmetrically, making clear that many of the things that characterize items also characterize persons, and vice versa. This is particularly relevant to tailored testing.

Conventional methods of tailored testing necessarily rely on accurate estimates of item difficulty and sicrimination. Giving items to large pretesting samples means that very often an item must be given to a person for

whom it is inappropriate. That is, there is an information function for items as well as persons, and when a high difficulty item is given to a low ability person, or vice versa, one learns little about either the item or the person.

In effect, the implied order system which is the basis for TAILOR tries to get around this by matching person to item as well as item to person. Consequently, the amount that it learns per response is likely to be greater than for other systems, provided the pretesting phase of the latter is included in the evaluation. Rather amazingly, it seems to be possible to make an appreciable savings on the basis of a rather small amount of data.

Applications

The effectiveness of TAILOR depends on having items which are of high discrimination. This tends to be true of all tailored testing schemes. If discriminations are only moderate, there is not a great deal more information provided by items near the person's ability level and those far from it. Moreover, it takes more items to focus accurately on the ability level.

Item parameters are always expressed relative to the standard deviation of ability. Thus, an item does not have an intrinsic discrimination; it depends on the population. Thus tailored testing is likely to be most applicable to situations where variability is greatest. This seems likely to be in placement situations and those in which training is being assessed. It is here that one is likely to find high variance, primarily as a function of whether training has been given or not. That is, there may well be a fairly definite joint order of persons and items.

TAILOR, or some descendant of it, may well be useful in such context,

particularly where large-scale testing is infeasible. This may well be true more generally if the strengths and weaknesses of it and other systems are realistically assessed. This will be particularly true if it is found, as seems likely, that persons behave systematically differently in computerized-tailored and conventional testing situations.

## Reference Note

1. Cudeck, R., Cliff, N., Reynolds, T. and McCormick, D. Monte Carlo results from a computer program for tailored testing. Technical Report No. 2, Department of Psychology, University of Southern California, 1976.

## References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (Part 5). Reading, Mass: Addison-Wesley, 1968.

Cliff, N. Complete orders from incomplete data: Interactive ordering and tailored testing. Psychological Bulletin, 1975, 82, 289-302.

Cudeck, R., Cliff, N., and Kehoe, J. TAILOR: A FORTRAN Program for interactive tailored testing. Educational and Psychological Measurement, In press.

Lord, F. M., and Novick, M. R. Statistical theories of mental test scores. Reading, Mass: Addison-Wesley, 1968, p. 112.

McCormick, D., and Cliff, N. TAILOR-APL: An interactive program for individual tailored testing. Educational and Psychological Measurement, In press.

McNemar, Q. Psychological Statistics (4th Ed.). New York: Wiley, 1969.

Nie, N., Null, C., Jenkins, J., Steinbrenner, K., and Bent, D. Statistical Package for the Social Science (2nd Ed.) New York: McGraw-Hill, 1975.

Urry, V. Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1974, 34, 253-269.

# Appendix

## Monte Carlo Data Generation

The data used during the Monte Carlo studies was produced at the beginning of each run in the following way. For simplicity only a general description is provided. Assume k persons and n items.

Step 1: Three vectors of length n and one of length k were produced which contained true scores for item discrimination, difficulty and guessing and the person ability score. These were normally distributed with a prescribed mean and standard deviation, given in:

$$c_i = SD\left[\sum_{j=1}^{12} UNI_j - 6.0\right] + \bar{X}$$

where

$c_i$ = the normally distributed item or person characteristic

SD = the standard deviation

UNI = a vector of uniform random numbers in the range of 0,1 based on a function available at USC's computer center

$\bar{X}$ = the desired mean

Step 2: A matrix $\underline{P}$ of rank k by n was computed, and contained the probabilities of a person answering an item correctly. As usual

$a_j$ = item discrimination

$b_j$ = item difficulty

$c_j$ = probability of chance success

$\theta_k$ = person ability

Appendix (Continued)

and

$$P_{ki} = c_j + (1 - c_j) \frac{1}{1 + \exp\left[-1.7a_j(\theta_k - b_j)\right]}$$

Step 3: A matrix $\underline{S}$ with rank k by n was generated, which is the score matrix based on $\underline{P}$ and a uniform random number where

$S_{ki} = 1$ if UNI $< P_{ki}$ ,

$S_{ki} = 0$ otherwise

and UNI is as given above, 1 and 0 indicate correct and incorrect responses.

Table A

Correlation Matrix of Principal Independent and Dependent Variables

from Monte Carlo Studies

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | Number of Persons | 1.000 | -0.258 | 0.032 | 0.248 | -0.077 | -0.059 | -0.134 | -0.101 | -0.089 | -0.281 |
| 2. | Number of Items | -0.258 | 1.000 | 0.164 | 0.135 | -0.258 | 0.333 | 0.403 | 0.305 | 0.286 | -0.572 |
| 3. | Mean Discrimination | 0.032 | 0.164 | 1.000 | 0.057 | 0.032 | 0.600 | 0.533 | 0.700 | 0.649 | -0.217 |
| 4. | Mean Chance | 0.248 | 0.135 | 0.057 | 1.000 | -0.138 | -0.175 | -0.209 | -0.246 | -0.222 | -0.049 |
| 5. | Mean Difficulty | -0.077 | -0.258 | 0.032 | -0.138 | 1.000 | 0.022 | -0.008 | 0.052 | 0.079 | 0.200 |
| 6. | Tau: Ability/Complete Test | -0.059 | 0.333 | 0.597 | -0.175 | 0.022 | 1.000 | 0.848 | 0.862 | 0.819 | -0.215 |
| 7. | R: Ability/Complete Test | -0.134 | 0.403 | 0.533 | -0.209 | -0.008 | 0.848 | 1.000 | 0.800 | 0.826 | -0.208 |
| 8. | Tau: Ability/Tailored Test | -0.101 | 0.305 | 0.700 | -0.246 | 0.052 | 0.862 | 0.800 | 1.000 | 0.924 | -0.221 |
| 9. | R: Ability/Tailored Test | -0.089 | 0.286 | 0.649 | -0.222 | 0.079 | 0.819 | 0.826 | 0.924 | 1.000 | -0.189 |
| 10. | Proportion of Relations | -0.281 | -0.572 | -.0217 | -0.049 | 0.200 | -0.215 | -0.208 | -0.221 | -0.189 | 1.000 |

Distribution List

Navy

4 Dr. Marshall J. Farr, Director
Personnel and Training Research Programs
Office of Naval Research (Code 458)
Arlington, VA 22217

1 ONR Branch Office
495 Summer Street
Boston, MA 02210
ATTN: Dr. James Lester

1 ONR Branch
1030 East Green Street
Pasadena, CA 91101
ATTN: Dr. Eugene Gloye

1 ONR Branch Office
536 South Clark Street
Chicago, IL 60605
ATTN: Dr. Charles E. Davis

1 Dr. M. A. Bertin
Scientific Director
Office of Naval Research
Scientific Liaison Group/Tokyo
American Embassy
APO San Francisco 96503

1 Office of Naval Research
Code 200
Arlington, VA 22217

6 Director
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Technical Director
Navy Personnel Research and Development Center
San Diego, CA 92152

1 Assistant for Research Liaison
Bureau of Naval Personnel (Pers Or)
Room 1416, Arlington Annex
Washington, CA 20370

1 Assistant Deputy Chief of Naval Personnel for Retension Analysis and Coordination (Pers 12)
Room 2403, Arlington Annex
Washington, DC 20370

1 CDR Paul D. Nelson, MSC, USN
Naval Medical R & D Command (Code 44)
National Naval Medical Center
Bethesda, MD 20014

1 Commanding Officer
Naval Health Research Center
San Diego, CA 92152
ATTN: Library

1 Chairman
Behavioral Science Department
Naval Command & Management Division
U. S. Naval Academy
Annapolis, MD 21402

1 Dr. Jack R. Borsting
U. S. Naval Postgraduate School
Department of Operations Research
Monterey, CA 93940

1 Director, Navy Occupational Task Analysis Program (NOTAP)
Navy Personnel Program Support Activity
Building 1304, Bolling AFB
Washington, DC 20336

1 Office of Civilian Manpower Management
Code 64
Washington, DC 20390
ATTN: Dr. Richard J. Niehaus

1 Superintendent
Naval Postgraduate School
Monterey, CA 93940
ATTN: Library (Code 2124)

1 Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054
ATTN: Dr. Norman J. Kerr

1 Principal Civilian Advisor
for Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508
ATTN: Dr. William L. Maloy

1 Director
Training Analysis & Evaluation Group
Code N-00t
Department of the Navy
Orlando, FL 32813
ATTN: Dr. Alfred F. Smode

1 Navy Personnel Research
and Development Center
Code 01
San Diego, CA 92152

5 Navy Personnel Research
and Development Center
Code 02
San Diego, CA 92152
ATTN: A.A. Sjoholm

2 Navy Personnel Research
and Development Center
Code 310
San Diego, CA 92152
ATTN: Dr. Martin F. Wiskoff

1 Navy Personnel Research
and Development Center
San Diego, CA 92152
ATTN: Library

1 Navy Personnel Research
and Development Center
Code 9041
San Diego, CA 92152
ATTN: Dr. J. D. Fletcher

1 D. M. Gragg, CAPT, MC, USN
Head, Educational Programs Development Department
Naval Health Sciences Education and Training Command
Bethesda, MD 20014

Army

1 Technical Director
U. S. Army Research Institute for the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Headquarters
U. S. Army Administration Center
Personnel Administration Combat Development Activity
ATCP- HRQ
Ft. Benjamin Harrison, IN 46249

1 Armed Forces Staff College
Norfolk, VA 23511
ATTN: Library

1 Dr. Stanley L. Cohen
U. S. Army Research Institute for the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Ralph Dusek
U. S. Army Research Institute for the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Joseph Ward
U. S. Army Research Institute for the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 HQ USAREUR & 7th Army
ODCSOPS
USAREUR Director of GED
APO New York 09403

1 ARI Field Unit - Leavenworth
Post Office Box 3122
Fort Leavenworth, KS 66027

1 Dr. Ralph Canter
U. S. Army Research Institute for the Behavioral and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Milton Maier
U. S. Army Research Institute
for the Behavioral and Social
Sciences
1300 Wilson Boulevard
Arlington, VA 22209

1 Dr. Milton S. Katz, Chief
Individual Training & Perfor-
mance Evaluation
U. S. Army Research Institute for
the Behavioral and Social
Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Air Force

1 Research Branch
AF/DPMYAR
Randoph AFB, Tx 78148

1 Dr. G. A. Echstrand (AFHRL/AST)
Wright Patterson AFB
Ohio 45433

1 AFHRL/DOJN
Stop #63
Lackland AFB, TX 78236

1 Dr. Martin Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

1 Dr. Alfred R. Fregly
AFOSR/NL
1400 Wilson Boulevard
Arlington, VA 22209

1 AFHRL/PED
Stop #63
Lackland AFB, TX 78236

1 Major Wayne S. Sellman
Chief of Personnel Testing
HQ USAF/DPMYP
Randolph AFB, Tx 78148

Marine Corps

1 Director, Office of Manpower
Utilization
Headquarters, Marine Corps (Code MPU)
MCB (Building 2009
Quantico, VA 22134

1 Dr. A. L. Slafkosky
Scientific Advisor (Code RD-1)
Headquarters, U. S. Marine Corps
Washington, DC 20380

1 Chief, Academic Department
Education Center
Marine Corps Development and
Education Command
Marine Corps Base
Quantico, VA 22134

1 Mr. E. A. Dover
2711 South Veitch Street
Arlington, VA 22206

Coast Guard

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (G-P-
1/62)
U. S. Coast Guard Headquarters
Washington, DC 20590

Other DOD

1 Dr. Harold F. O'Neil, Jr.
Advanced Research Projects Agency
Cybernetics Technology, Rm. 625
1400 Wilson Boulevard
Arlington, VA 22209

12 Defense Documentation Center
Cameron Station, Building 5
Alexandria, VA 22314
ATTN: TC

Other Government

1 Dr. Lorraine D. Eyde
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. William Gorham, Director
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. Vern Urry
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. Harold T. Yahr
Personnel Research and Development Center
U. S. Civil Service Commission
1900 E Street, N. W.
Washington, DC 20415

1 Dr. Andrew R. Molnar
Technical Innovations in Education Group
National Science Foundation
1800 G Street, N. W.
Washington, DC 20550

1 U. S. Civil Service Commission
Federal Office Building
Chicago Regional Staff Division
Regional Psychologist
230 South Dearborn Street
Chicago, IL 60604
ATTN: C. S. Winiewicz

1 Dr. Carl Frederiksen
Learning Division, Basic Skills Group
National Institute of Education
1200 19th Street, N. W.
Washington, DC 20208

Miscellaneous

1 Dr. Scarvia B. Anderson
Educational Testing Service
17 Executive Park Drive, N. E.
Atlanta, GA 30329

1 Mr. Samuel Ball
Educational Testing Service
Princeton, NJ 08540

1 Dr. Gerald V. Barrett
University of Akron
Department of Psychology
Akron, OH 44325

1 Dr. Kenneth E. Clark
University of Rochester
College of Arts and Sciences
River Campus Station
Rochester, NY 14627

1 Dr. John J. Collins
Vice President
Essex Corporation
6305 Caminito Estrellado
San Diego, CA 92120

1 Dr. Rene V. Dawis
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 Dr. Marvin D. Dunnette
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 ERIC
Processing and Reference Facility
4833 Rugby Avenue
Bethesda, MD 20014

1 Major I. N. Evonic
Canadian Forces Personnel Applied Research Unit
1107 Avenue Road
Toronto, Ontario, CANADA

1 Dr. Victor Fields
Montgomery College
Department of Psychology
Rockville, MD 20850

1 Dr. Edwin A. Fleishman
Visiting Professor
University of California
Graduate School of Administration
Irvine, CA 92664

1 Dr. John R. Frederiksen
Bolt, Beranek and Newman, Inc.
50 Moulton Street
Cambridge, MA 02138

1 Dr. Robert Glaser, Co-Director
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15213

1 Dr. Richard S. Hatch
Decision Systems Associates, Inc.
5640 Nicholson Lane
Rockville, MD 20852

1 Dr. M. D. Havron
Human Sciences Research, Inc.
7710 Old Spring House Road
West Gate Industrial Park
McLean, VA 22101

1 HumRRO Central Division
400 Plaza Building
Pace Boulevard at Fairfield Drive
Pensacola, FL 32505

1 HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921
ATTN: Library

1 Dr. David Klahr
Carnegie-Mellon University
Department of Psychology
Pittsburgh, PA 15213

1 Dr. Alma E. Lantz
University of Denver
Denver Research Institute
Industrial Economics Division
Denver, CO 80210

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
6780 Corton Drive
Santa Barbara Research Park
Goleta, CA 93017

1 Dr. William C. Mann
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina Del Rey, CA 90291

1 Mr. Edmond Marks
315 Old Main
Pennsylvania State University
University Park, PA 16802

1 Richard T. Mowday
College of Business Administration
University of Nebraska, Lincoln
Lincoln, NE 68588

1 Dr. Leo Munday, Vice-President
American College Testing Program
P. O. Box 168
Iowa City, IA 52240

1 Mr. Luigi Petrullo
2431 North Edgewood Street
Arlington, VA 22217

1 Dr. Steven M. Pine
University of Minnesota
Department of Psychology
Minneapolis, MN 55455

1 Dr. Diane M. Ramsey-Klee
R-K Research & System Design
3947 Ridgemont Drive
Malibu, CA 90265

1 Dr. Joseph W. Rigney
University of Southern California
Behavioral Technology Laboratories
3717 South Grand
Los Angeles, CA 90007

1 Dr. Andrew M. Rose
American Institutes for Research
3301 New Mexico Avenue, N. W.
Washington, DC 20016

1 Dr. George E. Rowland
Rowland and Company, Inc.
P. O. Box 61
Haddonfield, NJ 08033

1 Dr. Benjamin Schneider
University of Psychology
Department of Psychology
College Park, MD 20742

1 Dr. Lyle Schoenfeldt
Department of Psychology
University of Georgia
Athens, Georgia 30602

1 Dr. Arthur I. Siegel
Applied Psychological Services
404 East Lancaster Avenue
Wayne, PA 19087

1 Dr. Henry P. Sims, Jr.
Room 630 - Business
Indiana University
Bloomington, IN 47401

1 Dr. C. Harold Stone
1428 Virginia Avenue
Glendale, CA 91202

1 Dr. Patrick Suppes, Director
Institute for Mathematical Studies
in the Social Sciences
Stanford University
Stanford, CA 94305

1 Dr. Sigmund Tobias
PH.D Programs in Education
Graduate Center
City University of New York
33 West 42nd Street
New York, NY 10036

1 Dr. David J. Weiss
University of Minnesota
Department of Psychology
N660 Elliott Hall
Minneapolis, MN 55455

1 Dr. K. Wescourt
Stanford University
Institute for Mathematical Studies
in the Social Sciences
Stanford, CA 94305

1 Dr. Anita West
Denver Research Institute
University of Denver
Denver, CO 80210

1 Mr. George Wheaton
American Institutes for Research
3301 New Mexico Avenue, N.W.
Washington, DC 20016